Aligning Text-to-Image Models using Human Feedback

Year: 2023 | Citations: 368 | Authors: Kimin Lee, Hao Liu, M. Ryu, Olivia Watkins, Yuqing Du

---

Abstract

Deep generative models have shown impressive results in text-to-image synthesis. However, current text-to-image models often generate images that are inadequately aligned with text prompts. We propose a fine-tuning method for aligning such models using human feedback, comprising three stages. First, we collect human feedback assessing model output alignment from a set of diverse text prompts. We then use the human-labeled image-text dataset to train a reward function that predicts human feedback. Lastly, the text-to-image model is fine-tuned by maximizing reward-weighted likelihood to improve image-text alignment. Our method generates objects with specified colors, counts and backgrounds more accurately than the pre-trained model. We also analyze several design choices and find that careful investigations on such design choices are important in balancing the alignment-fidelity tradeoffs. Our results demonstrate the potential for learning from human feedback to significantly improve text-to-image models.